DISTRIBUTED LOAD BALANCING FOR SINGLE ENTRY-POINT SYSTEMS
Abstract of the Disclosure

A method and system for distributing work load in a cluster of at least two service resources. Depending upon the configuration, a service resource may be an individual process, such as a single instance of a computer game, or a node on which multiple processes are executing, such as a Server. Initial connection requests from new clients are directed to a single entry-point service resource in the cluster, called an intake. A separate intake is designated for each type of service provided by the cluster. The clients are processed in a group at the service resource currently designated as the intake to which clients initially connected, for the duration of the session. Based upon its loading, the current intake service resource determines that another service resource in the cluster should become a new intake for subsequent connection requests received from new clients. Selection of another service resource to become the new intake is based on the current work load of each resource in the cluster. All resources in the cluster are periodically informed of the resource for each service being provided that was last designated as the intake, and of the current load on each resource in the cluster. Subsequently, new clients requesting a service are directed to the newly designated intake for that service and processed on that resource for the duration of the session by those clients.